NUCLEIC ACID ANALYSIS METHOD AND SYSTEM

FIELD OF THE INVENTION

The present invention is in the field of biochemical assays and more specifically biochemical assays for the determination and identification of nucleic acids.

BACKGROUND OF THE INVENTION

There are many instances when it is desirable to detect or identify nucleic acids in a sample. For example, many disease states are characterized by differences in the expression levels of one or more genes. Thus, altered expression of oncogenes or tumor suppressor genes leads to cancer, while viral infection is characterized by the expression of viral genes in a host cell. Since the expression level of a particular gene in a cell is usually proportional to the amount of mRNA transcribed from the gene, malignant transformation or viral infection is detected by determining the amount of mRNA for a relevant gene in the cell and comparing it with known controls.

Blotting techniques have frequently been used to identify nucleic acids in a mixture of oligonucleotides. The mixture is first fractionated by gel electrophoresis, and the separated oligonucleotides are then blotted from the gel onto a nitrocellulose sheet. The sheet is then incubated in the presence of one or more labeled DNA probes having complementary nucleotide sequences to the oligonucleotides of interest on the blot (referred to as target oligonucleotides). A target oligonucleotide is then detected following its hybridization to its labeled probe. In practice, however, these methods suffer from several disadvantages. Two

or more oligonucleotides of similar molecular weights may not be resolved by the electrophoresis. Moreover, low hybridization efficiency and cross reactivity of the probes make it difficult to obtain an accurate quantitative measure of the amount of a target present in the original mixture.

Another approach to detecting and identifying oligonucleotides in a mixture uses oligonucleotide probes immobilized on a solid support. Such probe arrays are synthesized using methods of spatially addressed parallel synthesis in which many oligonucleotide probes are simultaneously synthesized in a highly parallel fashion while attached at one end to the support surface. The solid support may have a very small surface area (typically about 1-2 cm^2) while comprising over 1,000,000 different oligonucleotide probes. The probes typically have lengths of 20 to 25 nucleotides, and the location of each different oligonucleotide probe in the array is known. The bases in the oligonucleotide probes may be joined by a linkage other than a phosphodiester bond, so long as it does not interfere with hybridization. In particular, oligonucleotides may be peptide nucleic acids in which the constituent bases are joined by peptide bonds rather than phosphodiester linkages. The probes may be attached to the support either directly or indirectly by covalent binding, hydrogen bonding, ionic interaction, hydrophobic interaction, etc. An oligonucleotide may include natural (i.e. A, G, C, U or T) or modified bases (7-deazaguanosine, inosine, etc.). The high-density array may also contain a number of control probes such as normalization controls, expression level controls, and mismatch controls.

Methods of forming such high-density probe arrays with a minimal number of synthesis steps are known. A probe array can be synthesized on a solid substrate by a variety of methods, such as light-directed chemical coupling, and mechanically directed coupling, as disclosed in Pirung *et al.* U.S. Patent No. 5,143,854, and PCT Publication Nos. WO 92/10091, WO 93/98668 and WO 90/15070. PCT Publication No. WO/97/27317 discloses use of such arrays to detect targets comprising a specific nucleotide sequences.

SUB A1

High-density arrays are suitable for the quantification of small variations in the abundance of a target of interest present in a target mixture at a concentration as low as 1 per 1,000,000 oligonucleotides. A labeled target mixture may be used containing, for example, mRNA transcripts whose concentrations in the target mixture are proportional to the expression level of the genes in the cells in which they were transcribed. Frequently the oligonucleotides in the target mixture are amplified prior to performing the assay by quantitative PCR or reverse transcriptase PCR.

The probe array is incubated in the presence of the labeled targets. If a probe for a target is present in the array, the two will stably hybridize while other targets in the mixture will not. After the incubation, unbound targets are removed. The array is then examined for the presence and location of label in the array. If, for example, the targets have been fluorescently labeled, each location in the array is individually exited at the excitation wavelength of the label, and the fluorescent emission intensity at each location is measured. This is most conveniently accomplished using a confocal microscope automated with a computer-controlled stage which automatically scans the entire probe array. The microscope may be equipped with a phototransducer attached to an automated data acquisition system to automatically record the fluorescence signal at each location in the array. The signal intensity associated with each probe species in the array is proportional to the number of targets hybridized to the probe. Effective detection and quantitation of hybridization typically requires about 20 copies of each probe species.

Where the concentrations of the nucleic acids comprising the samples reflects transcription levels of genes in a cell from which the targets were derived, these screening methods permit identification of differences in transcription (and by implication in expression) of the nucleic acids comprising the two or more samples. The labeling pattern in the array thus forms a *"fingerprint"* of the gene expression in the cell. Such fingerprints can be used, for example, to distinguish normal and abnormal cells.

The generic difference screening methods are advantageous in that they require no *a priori* assumptions about the sequences of oligonucleotides in the probe array. The sequences of the probe oligonucleotides may even be an arbitrarily selected subset of an oligonucleotide probe family. Even in these cases, since the sequence of each probe in the array is known, generic difference screening provides direct sequence information regarding the differentially expressed nucleic acids in the sample.

"*Expression monitoring*" is used to determine absolute levels of targets in a target mixture. A high density probe array is prepared wherein the probes are selected to be complementary to subsequences of the targets of interest in the target mixture. If a probe species is present in the array in excess in comparison to the number of copies of its complementary target in the target mixture, an essentially accurate absolute measurement of the expression level of the genes of interest is obtained.

Unlike generic screening methods, in expression monitoring, the probe array must contain only probes, each of which hybridizes specifically to a single, predetermined target of interest with no non-specific binding or cross-hybridization. This places a major obstacle to the application of expression monitoring because probes often cross hybridize with several targets due to the presence of complementary subsequences in several targets. Furthermore, hybridization between a probe and a target may occur in spite of minor mismatches between the two. Therefore, probes that show poor specificity to a target mixture of interest must first be identified, and excluded from the probe array. Since the number of probe species in the probe array must be equal to the number of target species in the target mixture, very large probe arrays are needed to analyze complex target mixtures. Moreover, because the observed hybridization of a probe to a target is prone to high variability due to reaction condition and measurement "*noise*". in practice, chip designs typically include about twenty specific different oligonucleotide probes, in addition to control probes, for each target of interest. As chip space is limited and the number of targets to be analyzed increases, there is a

considerable need for methods for unambiguously detecting a target with as few probes as possible so as to increase the number of targets that can be detected with a single chip.

GLOSSARY

The following is a glossary of some terms used in the present specification:

The term ***"probe oligonucleotide"*** (at times also ***"probe"***), used above and further below means to denote an oligonucleotide, typically immobilized on a substrate, which comprises a probing nucleotide sequence and possibly non-probing nucleotide sequences, such as a non-probing sequence by which the probe oligonucleotide is immobilized on the substrate. A probe oligonucleotide may at times consist entirely of probing nucleotide sequences. The probe oligonucleotide may be DNA, RNA, PNA, or generally nucleotides connected to one another by any suitable backbone which does not interfere or which minimally interferes with the ability of the oligonucleotide to hybridize with essentially complementary sequences. The term ***"probing nucleotide sequence"*** or ***"probing sequence"*** refers to a sequence contained within a probe oligonucleotide which can hybridize (and bind to) with an essentially complementary sequence in a target oligonucleotide. It should be noted that the probing sequence is not necessarily a single contiguous region within the probe oligonucleotide. Thus, the probing sequence may consist of any number of contiguous regions separated by non-probing sequences of the probe oligonucleotide.

The term ***"probing unit"*** means to denote a group of probe oligonucleotides, which are defined in terms of their location and target specificity. The probing unit may consist of one species of probe oligonucleotide with one or more probing nucleotide sequences which can hybridize to one or more target oligonucleotides; or may consist of a number of probe oligonucleotide specie, each with its one or more probing nucleotide sequences, which in combination can hybridize to one or more target oligonucleotides. Each probing unit is a defined in

terms of the target oligonucleotides that can hybridize thereto. The target specificity of each probing unit is determined by its one or more probing nucleotide sequences: in case the probing unit consists of more than one probe oligonucleotide each with a different target specificity, the target specificity will be a combination of the target specificities of the different probe oligonucleotides (in carrying out the assay in accordance with the invention it is not possible to know which of the different probe oligonucleotides hybridized to a target). The probing units are typically contained all on a single substrate, although it is possible to include them on a number of different substrates as long as the location of each probing unit is known. A typical carrying substrate is that known in the art as a "chip" or "wafer". On such a chip the probing units are arranged in an array with the location of each probing unit being defined and known.

The term ***"target oligonucleotides"*** (at times also ***"target"***) means to denote an oligonucleotide to be assayed in a sample. This is typically an mRNA or a cDNA derived therefrom. The term ***"target nucleotide sequence"*** or ***"target sequence"*** refers to a sequence within the target oligonucleotide which hybridizes to the probing nucleotide sequence. It should be noted that a target oligonucleotide may at times comprise two or more different target nucleotide sequences, each one being essentially complementary to a different probing nucleotide sequence. The target oligonucleotide can be derived from essentially any source of nucleic acids (e.g., including, but not limited to chemical syntheses, amplification reactions, forensic samples, etc.) It is either the presence or absence of one or more target oligonucleotide that is to be detected, or the amount of one or more target oligonucleotide that is to be quantified. The target oligonucleotide(s) that are detected preferentially have nucleotide sequences that are complementary to the nucleic acid sequences of the corresponding probe oligonucleotide(s) to which they specifically bind (hybridize).

The term ***"essentially complementary"*** or the term ***"complementary"*** used above and below means to denote that the target sequence and the probing sequence have a degree of complementarity that allows them to hybridize under

appropriate conditions. At times two essentially complementary sequences may be 100% complementary. although at other times the degree of complementarity may be less. e.g. 90% or at times even 80%. It should be noted that some degree of hybridization may result also in case where the complementarity is less than 100%. Obviously. the hybridization affinity is less in case of a non perfect complementarity (complementarity of less than 100%). When the probing sequence consists of more than one contiguous region within the probe oligonucleotide, binding of a target to the probing sequence may give rise to hairpin-like structures in the probe oligonucleotide or in the target oligonucleotide, in which an unhybridized, non-probing, sequence intervenes between two regions in the probing sequence which have hybridized to the target. The term "(essentially) complantary" relates also to such a scenario.

The term ***"sample"*** denotes a medium, usually liquid, presumed to contain targets of interest. The sample may also be a processed original sample or a fraction from an original sample which contains its oligonucleotides.

The term ***"species"*** denotes one specific probe or target. One probe or target species differs from another by at least one nucleotide.

The term ***"array of probes"*** denotes a predetermined spatial arrangement of probe species present on a *solid substrate* (see below) or on a *multi-well arrangement* (see below), where all probes of the same species are confined to a separate, specific, and known location in the array.

The term ***"location"*** or ***"Coordinate"*** denotes one specific distinct area in the array holding one or more known species of probes.

The term ***"chip"*** denotes an array present on a solid substrate.

The term ***"solid substrate"*** denotes a rigid or semi-rigid surface on which the probes are immobilized. Immobilization may be directly to the surface or indirectly through linking moieties and includes attachment by covalent binding, hydrogen binding, ionic interactions, hydrophobic interactions and the like.

The term ***"multi-well arrangement"*** denotes a device having a plurality of liquid-holding wells where each well is in fact a *"location"* of the array.

The term ***"determination"*** or ***"determining"*** denotes either a qualitative determination of the presence of a certain target oligonucleotide in a sample, or, by some embodiments of the invention. a quantitative determination of the level of the target oligonucleotide in the sample. A quantitative determination is typically a determination of the relative abundance of the target oligonucleotide in the sample as compared to other target oligonucleotides. In a biological sample obtained from a certain tissue, such a quantitative determination may give an indication of the relative expression level of the target oligonucleotide in such a sample.

SUMMARY OF THE INVENTION

Probe oligonucleotide arrays hitherto used were designed subject to various constraints. Constraints on prior art arrays include fixed probe length, probe physicochemical properties (affinity of hybridization to complementary sequences) and the requirements for target specificity of different probe species, and others. Target specificity means that the assayed target sequence of a probe appears in only one target species in the sample. In accordance with the invention, the probe array is released from the constraint of specificity so that (i) an oligonucleotide probe species in a location of a probe array may have a probing sequence capable of hybridizing with more than one target oligonucleotide or (ii) oligonucleotide probes in a single location may be composed of oligonucleotides of a number of different specie such that the ensemble of oligonucleotides in each location can bind to more than one target oligonucleotide. This allows the probe array to be designed with substantially less than 20 probe locations for each target species, for probes of a length of about 25 nucleotides.

In accordance with a first aspect of the invention there is provided an ensemble of **k** different probing units, for determining, by hybridization, **n** different target oligonucleotides in an assayed sample; each of said probing units comprises one or more probe oligonucleotides with one or more probing nucleotide sequences and each of said target oligonucleotides comprising one or more target nucleotide sequences. with the probing nucleotide sequences being capable of hybridizing to

target nucleotide sequences, characterized in that the probing nucleotide sequences of at least one probing unit can hybridize to target nucleotide sequences in at least two different target oligonucleotides.

In accordance with another aspect, the present invention provides a method for designing a system for determining **n** target oligonucleotides, S_1, S_2, ..., S_n, in a sample, comprising:

(a) selecting or designing an ensemble of **k** probing units, P_1, P_2, ..., P_k, each probing unit consisting of one or more probe oligonucleotide species having in combination one or more probing nucleotide sequences, the one or more probing nucleotide sequences of at least one of the probing units can hybridize to target nucleotide sequences in at least two different target oligonucleotides;

(b) arranging the ensemble of said probing units in a manner allowing exposure to the sample under conditions permitting hybridization between corresponding target oligonucleotide and probe oligonucleotide sequences and allowing determination of an hybridization event and the extent of hybridization for each of the probe oligonucleotides;

(c) devising **T** being an **k** × **n** mathematical matrix consisting of components t_{ij}, in which matrix each t_{ij} denotes the affinity of hybridization of a target oligonucleotide S_i to probe oligonucleotides of probing unit P_j, under defined assay conditions (namely conditions to be eventually applied in the assay – type of medium, its content, temperature, etc.); and

(d) designating the **T** matrix as being associated with said ensemble to permit its use in determining expression of each of said target oligonucleotides.

By a further aspect, the present invention provides a method for determining relative abundance of **n** target oligonucleotides S_1, S_2,....., S_n, in an assayed sample, comprising:

(a) providing an ensemble of **k** probing units, P_1, P_2, ..., P_k, each probing unit consisting of one or more probe oligonucleotide species having in combination one or more probing nucleotide sequences, the probing units being selected such that at least one has aof the probing nucleotide sequences of at least one probing unit can hybridize to target nucleotide sequences in at least two different target oligonucleotides;

(b) exposing said ensemble to the assayed sample under hybridization-permissive conditions and measuring level of hybridization of target oligonucleotides from the assayed sample to each of the probing units;

(c) in a processor, devising a k-dimensional vector $c = (c_1, ..., c_k)$, consisting of k coordinates c_j, with j being an integer from 1 to k, each of coordinates c_j being either (i) a representation of the level of target oligonucleotides hybridized to probing unit P_j, or (ii) a representation of the difference between said level and a level measured in an identical ensemble exposed to a control sample in the same manner to that defined in step (b) (in the latter case the vector c is in fact a product of subtraction of two vectors consisting each of results obtained from a different sample);

(d) in the processor, calculating an n-dimensional vector e, consisting of n coordinates e_i, each of coordinates e_i being an indication of the level of target S_i in the sample, by solving the following vector equation (1):

$$\mathbf{c = Te} \qquad (1)$$

in which **T** is a $\mathbf{k \times n}$ mathematical matrix consisting of components t_{ij}, in which matrix each each t_{ij} denotes the affinity of hybridization of a target oligonucleotide S_i to probe oligonucleotide P_j under the assay conditions.

In accordance with one embodiment of the invention, particularly where the vector $c = (c_1, \ldots, c_k)$ is that defined under (i) in step (c), the method comprises an additional step in which the calculated vector e is subtracted from another vector e_c, which is obtained in the same manner with a control sample, to obtain a vector e_s, which consists in fact of the values for expression of the target oligonucleotides which are either (i) expressed only in the assayed sample and not in the control sample, (ii) expressed only in the control sample and not in the assayed sample, or (iii) expressed at a different level in the assayed and in the control level. In the following when discussing vector e, it is meant to refer, *mutatis mutandis*, also to vector e_s.

The invention provides, by a still further aspect, a system for determining relative abundance of **n** target oligonucleotides $S_1, S_2, \ldots, S_n$, in an assayed sample, comprising:

(i) an ensemble of **k** probing units, $P_1, P_2, \ldots, P_k$, each probing unit consisting of one or more probe oligonucleotide species having in combination one or more probing nucleotide sequences, the probing units being selected such that at least one of the probing nucleotide sequences of at least one probing unit can hybridize to target nucleotide sequences in at least two different target oligonucleotides;

(ii) detector for detecting a quantity indicating hybridization of a target oligonucleotide to a probing unit;

(iii) a processor coupled to said detector for constructing, based on the detected quantity, a k-dimensional vector $c = (c_1, \ldots, c_k)$, consisting of k coordinates c_j, with j being an integer from 1 to k, each of coordinates c_j being either (i) a representation of the level of target oligonucleotides hybridized to probing unit P_j, or (ii) a representation of the difference between said level and a level measured in an identical ensemble exposed to a control sample in the same manner to that defined in step (b); and for calculating an n-dimensional vector e, consisting of n coordinates e_i, each of

coordinates e_i being an indication of the level of target S_i in the sample. by solving the following vector equation (1):

$$\mathbf{c} = \mathbf{Te} \qquad (1)$$

in which **T** is a $\mathbf{k} \times \mathbf{n}$ mathematical matrix consisting of components t_{ij}, in which matrix each each t_{ij} denotes the affinity of hybridization of a target oligonucleotide S_i to probe oligonucleotide P_j under the assay conditions.

The detected quantity may be a fluorescent label, a radioactive label, etc., as known *per se*. In general any signal which can be used to detect the occurrence of hybridization may be such a detectable quantity. The detectable quantity may also at times be the disappearance of a signal, e.g. as a result of dissociation of a labeled oligonucleotide from the probe oligonucleotide in the presence of the target, etc.

The invention provides, by yet another aspect, for use in an assay for determining relative abundance of **n** target oligonucleotides S_1, S_2,....., S_n, in an assayed sample, a combination comprising:

(i) an ensemble of **k** probing units, P_1, P_2, ..., P_k, each probing unit consisting of one or more probe oligonucleotide species having in combination one or more probing nucleotide sequences, the probing units being selected such that at least one of the probing nucleotide sequences of at least one probing unit can hybridize to a target nucleotide sequences in at least two different target oligonucleotides;

(ii) a computer readable medium, which may be a magnetic disk, a CD-ROM or any other suitable computer readable medium. carrying data for inputting to a processor, which processor, based on an inputted data constructs a vector $c = (c_1, \ldots, c_k)$, consisting of k coordinates c_j, with j being an integer from 1 to k, each of coordinates c_j being either (i) a value representing the level of target oligonucleotides hybridized to probing unit P_j, or (ii) a value representing the difference between said level and a level measured

in an identical ensemble exposed to a control sample in the same manner to that defined in step (b); calculates an n-dimensional vector c. consisting of n coordinates e_i, each of coordinates e_i being an indication of the level of target S_i in the sample, by solving the following vector equation (1):

$$\mathbf{c} = \mathbf{Te} \qquad (1)$$

in which **T** is a $\mathbf{k} \times \mathbf{n}$ mathematical matrix consisting of components t_{ij}. in which matrix each t_{ij} denotes the affinity of hybridization of a target oligonucleotide S_i to probe oligonucleotide P_j under the assay conditions; said data on said data carrier comprises said matrix T which is associated for use with said ensemble.

The ensemble is preferably in a form of an oligonucleotide chip/wafer. The probe oligonucleotides are typically DNA oligonucleotides (consisting of deoxy ribonucleotides). On such a wafer each probing unit is immobilized at a distinct location. Thus. a target oligonucleotide detected at a specific location can be associated with a specific probing unit to which it hybridized.

The determination of the target in a sample, in accordance with the invention. may either be a qualitative determination of presence or absence of the target oligonucleotide in the sample, or may be a quantitative determination of its relative abundance. In a qualitative determination in accordance with the invention, a certain threshold level will be given for the values of the coordinates of vector e, and each value above the threshold will be interpreted as presence of a certain target oligonucleotide P_i in the sample; and a value below the threshold will be considered as absence of target S_i from the sample (the terms "above" and "below" should be understood in their absolute sense as in the case of the vector e_s, its coordinates may also assume negative values. In a quantitative determination, each of the vector's coordinates will be regarded as representing relative abundance of the corresponding target S_i in a sample.

Matrix **T**. in accordance with one, simplified, embodiment of the invention, consists of values 1 and 0: wherein the value of $t_{ij} = 1$ represents the ability of target

oligonucleotide S_j to hybridize to probe oligonucleotide P_i; whereas the value of $t_{ij} = 0$ will signify the lack of ability of S_j to hybridize to P_i.

In accordance with another embodiment of the invention, each of values t_{ij} is a non-negative value representing the relative affinity of hybridization of S_i to P_j under the assay conditions.

Vector equation (1) defines a system of k linear equations in n unknown. In accordance with the invention, a selection of probes may be made when this system of equations is overdetermined. This could be the case, for example, when n>k. In this case, solving the vector equation (1) involves finding a vector **e** that meets a predetermined criterion. In one embodiment, an error vector $d = (d_1,\ldots,d_k)$ is defined, for example, by

$$(d_1,.....d_k) = \sum_{i=1}^{k}(c_i - \sum_{j=1}^{n} t_{ij}e_j)^2 \qquad (2)$$

and the predetermined criterion which **e** must meet is that it minimizes the error function. It should be note that equation (2) is but an example and many other equations for calculating an error vector may be employed. Other criteria for defining a solution in the case that the system of equations is overdetermined are also envisaged within the scope of the invention.

One possible, but not exclusive example of an application, is in cases in which it is *a priori* known that the set of possible vectors **e** is finite. This would be the case, for example, when it may be presumed that the components of **e** are small integral multiples of a fundamental signal unit. The set of possible vectors **e** may, by another example be restricted if it is *a priori* known that no more than a pre-specified number of components in each possible vector **e** are non-zero or by looking at a solution in which the number of non-zeros is minimal. This would typically be the case, for example, in a differential assay in which the components of the vector **e** represent the difference in expression levels of targets in two

different target samples. In cases such as these, in accordance with the invention, a selection of probes may be made for which the ensuing system of k linear equations in n unknowns defined by vector equation (1) is underdetermined. At times the accurate solution is not necessary and it is enough to obtain an approximate solution with accuracy sufficient to be able to identify a change versus a control or reference sample. In other words, the result of a certain vector e_1 may have a certain degree of uncertainty such, however that would permit to differentiate it from a vector e_2 obtained with a reference sample. The degree of permissible uncertainty depends on the nature of the performed assay.

Each probing sequence is essentially complementary to at least one target sequence. The case where the probing sequence is complementary by 90-100%, is a preferred embodiment in accordance with the invention. The probing sequence will usually have less than 5, preferably less than 3 and typically between 0 to 2 mismatches.

The probing sequence of the probe oligonucleotides may have a length of about 12 to about 80 mer, typically less than about 70 mer and preferably within the range of about 15 to about 60 mer.

Preferably, a plurality of probing units in the ensemble has probing sequences that allow them to hybridize to two or more target nucleotide sequences. A probing nucleotide sequence that can hybridize to more than when target may, for example be in the case of targets which are alternative splicing variants of the same gene.

Usually, the ensemble of probing units with their probing sequences is such that a single target oligonucleotide can hybridize to more than one probe oligonucleotide. The probe oligonucleotides can thus be regarded as belonging to groups, with all oligonucleotide probes of the same group having the common feature that they can hybridize to a target sequence in one target oligonucleotide. As it will be appreciated, in view of the lack of specificity in the probing sequences of at least one probe, a single probe oligonucleotide can belong to more than one group: in other words, different groups share probe oligonucleotides between them.

This is in marked distinction to the prior art which required that each group of probe oligonucleotides consists of oligonucleotides binding the same target; namely in the prior art. no probing oligonucleotides were shared between different groups. This feature of the invention allows the use of a marked lower amount of probe oligonucleotide needed in order to test a given amount of target oligonucleotides. Notwithstanding the lower amount of the probes, the fact that a target oligonucleotide can hybridize to a plurality of probe oligonucleotides, permits a reduction in the "*noise*" of the assay. It should however be noted that in order to allow a meaningful result. the number of probe oligonucleotides which are shared between two groups is less than the number of probe oligonucleotides of at least one of the two groups.

A result of the invention is thus a reduction in the number of probe oligonucleotides **k**, required to assay a set of target oligonucleotides **n**. While in the prior art, where the probe oligonucleotides have the characteristic length of about 25 mer, **k** typically equals about 20n. In accordance with the invention, **k** is typically less than about 10n, preferably less than about 4n, most preferably less than about 2n and at times about equal to or at times even less than about **n**.

The probing units are usually designed such that each target oligonucleotide will hybridize to only a few of the probe oligonucleotides. (See below regarding the "*sparse vector*"). This permits a high degree of accuracy when assaying target oligonucleotides in a sample with the ensemble of the invention.

The ensemble of probing units may be provided in any suitable form that permits hybridization of matching target oligonucleotides (namely target oligonucleotides with a target sequence which is complementary or essentially complementary to the probing sequence). Another requirement is the ability to identify the occurrence of hybridization of target oligonucleotides to each probing unit. Typically. therefore. the ensemble is immobilized on a substrate, which may be a micro-well array or, preferably, a solid substrate, known in the art as a "*chip*". The array is produced such that each probe oligonucleotide is at a defined location on the substrate.

The matrix **T** may be determined empirically by exposing each probing unit individually to each target oligonucleotide, under normalized conditions and determining the degree of hybridization of the target oligonucleotide to the probe oligonucleotide. In accordance with another embodiment, the matrix **T** may be determined from theoretical considerations of the expected hybridization affinity of each target oligonucleotide to each of the probe oligonucleotides. For example, this may be based on the chemical properties, e.g. the ratio of G and C content to the A and T content of the target and the probe oligonucleotide sequences.

The probe oligonucleotides are typically selected using optimization models using computer simulations. For example, in the case of a group of known targets (as will be appreciated not always all targets be known in view of potential existence of unknown splice variants of some of the targets) there is a finite list of potential probing sequences. As an example, for 1000 targets, each having 500 nucleotide bases, there are about 950,000 potential probing sequences of 50mer (about 950 for each target) (in fact the number may be somewhat larger since also certain non-perfectly matched probing sequences can be used as they may also hybridize to target oligonucleotides. From these potential probes a certain arbitrary number of probes may initially be chosen and checked by a computerized simulated assay with different simulated expression patterns of target oligonucleotides and the results are scored.

The scoring may be performed by first determining binding of the simulated expressed targets to each chosen probing sequence, at times determining it qualitatively ("+" or "-") or quantitatively (degree of binding), and then solving to determine a vector c. Then the vectorial equation $c = Te$ is solved to find e and then the difference between the calculated e vector (e_{cal}) and the simulated vector (e_{sim}) (the simulated expression level of the different targets) is used to score the simulated result. The simulated assay may be repeated a number of times, each time using a different e_{sim} and a new score based on the e_{sim} and e_{cal} difference is obtained. Eventually a range of scores is obtained. The list of chosen simulated probes may then be modified a new range of scores obtained until and this

simulated ensemble choosing process may then continue until an optimal choice of probes is achieved.

In the assay the probe array is incubated in the presence of labeled target mixture under conditions permitting the hybridization of each labeled target species to its several matching probe specie. Unbound targets are removed and the amount of label associated with each probe species in the probe array is measured.

c_i, (i=1,...,k) is the amount of measured label observed associated with probe species i. e_j, (j=1,...,n) is the relative abundance of target species j in the original target mixture. The values of the e_j are related to the measured label values c_j by the matrix equation:

$$\mathbf{c} = \mathbf{Te} \qquad (1)$$

c is the k-dimensional column vector which can be measured. The n-dimensional column vector **e** of the different probe species in the original target mixture is an unknown to be determined. Equation **(1)** thus defines a system of k linear equations in n unknowns.

The invention will now be illustrated in the following more Detailed Description of the Invention, which should not be construed as limiting.

BRIEF DESCRIPTION OF THE DRAWINGS

Figs. 1-3 illustrate, schematically, a manner of carrying out an assay in accordance with the invention.

DETAILED DESCRIPTION OF THE INVENTION

I. ARRAYS OF NUCLEIC ACIDS ON SOLID SUBSTRATES

A. Substrates

The term *"substrate"* refers to a material having a rigid or semi-rigid surface. In many cases, at least one surface of the substrate will be substantially flat

or planar. although in some cases it may be desirable to physically separate synthesis regions for different nucleic acids with, for example, wells, raised regions. etched trenches. or the like. According to other embodiment, small beads may be provided on the surface which may be released upon completion of the synthesis. Preferred substrates generally comprise planar crystalline substrates such as silica based substrates (e.g. glass, quartz, or the like), or crystalline substrates used in. e.g., the semiconductor and microprocessor industries, such as silicon, gallium arsenide and the like. These substrates are generally resistant to the variety of synthesis and analysis conditions to which they may be subjected. Particularly preferred substrates will be transparent to allow the photolithographic exposure of the substrate from either direction.

Silica aerogels may also be used as substrattes. Aerogel substrates may be used as freestanding substrates or as a surface coating for another rigid substrate support. Aerogel substrates provide the advantage of large surface area for nucleic synthesis. e.g., 400 to 1000 cm^2/gm, or a total useful surface area of 100 to 1000 cm^2 for a 1 cm^2 piece of aerogel substrate. Such aerogel substrates may generally be prepared by methods known in the art, e.g., the base catalyzed polymerization of $(MeO)_4Si$ or $(EtO)_4Si$ in ethanol/water solution at room temperature. Porosity may be adjusted by altering reaction condition by methods known in the art.

Individual planar substrates generally exist as wafers which can have varied dimensions. The term *"wafer"* generally refers to a substantially flat sample of substrate from which a plurality of individual arrays or chips may be fabricated.

The size of the substrate wafer is generally defined by the number and nature of arrays that will be produced from the wafer.

Although primarily described in terms of flat or planar substrates, the present invention may also be practiced with substrates having substantially different conformations. For example, the substrate may exist as particles, strands, precipitates. gels. sheets. tubing, spheres, containers, capillaries, pads. slices, films, plates. slides. etc. In a preferred alternate embodiment, the substrate is a glass tube

or microcapillary. The capillary substrate provides advantages of higher surface area to volume ratios, reducing the amount of reagents necessary for synthesis. Similarly, the higher surface to volume ratio of these capillary substrates imparts more efficient thermal transfer properties.

B. Synthesis of nucleic acid arrays

General methods for the solid phase synthesis of a variety of polymer types have been previously described. Methods of synthesizing arrays of large numbers of polymer sequences, including oligonucleotides and peptides, on a single substrate have also been described. See U.S. Patent Nos. 5,143,854 and 5,384,261 and Published PCT Application No. WO 92/10092, each of which is incorporated herein by reference in its entirety for all purposes.

As described previously, the synthesis of oligonucleotides on the surface of a substrate may be carried out using light directed methods as described in, e.g. U.S. Patent Nos. 5,143,854 and 5,384,261 and Published PCT Application No. WO 92/10092, or mechanical synthesis methods as described in 5,384,261 and published PCT Application No. 93/09668, each of which is incorporated hereby by reference. Preferably, synthesis is carried out using light-directed synthesis methods. In particular, these light-directed or photolithographic synthesis methods involve a photolysis step and a chemistry step. The substrate surface, prepared as described in the publication comprise functional groups on its surface. These function groups are protected by photolabile protecting groups (*"photoprotected"*). During the photolysis step, portions of the surface of the substrate are exposed to light or other activators to activate the functional groups within those portions, i.e., to remove photoprotecting groups. The substrate is then subjected to a chemistry step in which chemical monomers that are photoprotected at at least one functional group are then contacted with the surface of the substrate. These monomers bind to the activated portion of the substrate through an unprotected functional group.

Subsequent activation and coupling steps couple monomers to other preselected regions, which may overlap with all or part of the first region. The activation and coupling sequence at each region on the substrate determines the

sequence of the polymer synthesized thereon. In particular, light is shown through the photolithographic masks which are designed and selected to expose and thereby activate a first particular preselected portion of the substrate. Monomers are then coupled to all or part of this portion of the substrate. The masks used and monomers coupled in each step can be selected to produce arrays of polymers having a range of desired sequences, each sequence being coupled to a distinct spatial location on the substrate which location also dictates the polymer's sequence. The photolysis steps and chemistry steps are repeated until the desired sequences have been synthesized upon the surface of the substrate.

By another synthesis method DNA oligonucleotides are attached to glass slides (Southern, E.M. *Nuc. Acids. Res.*, **22**:1368-1373, 1994). In subsequent synthetic steps, these oligonucleotides are elongated by presenting nucleotides to defined areas on the slides. After the synthesis is complete, labeled complementary probes are hybridized to the target DNA on the slide. Similarly, arrays of DNA probes can be synthesized on aminated polypropylene film using a controlled photodeprotection chemistry and photoprotected N-acyl-deoxy-nucleoside phosphoramidites (Matson, R., *Anal. Biochem.*, **224:**110-116, 1995). Methods which do not include direct synthesis on the support can also be used which involve the attachment of PCR products to silylated glass slides (Schena, M., *PNAS*, **93:**10614-10619, 1996).

The probes may be arranged in any desired array on the solid substrate using a variety of techniques including: use of light to direct the combinatorial chemical synthesis of biopolymers on a solid support; embedding of DNA sequences on a gel coated chip (Edginton *Bio/Technology*, **12**:468-471, 1994; Yershov *et al.*, *PNAS*, **93**:4913-4918, 1996); micropatterning lipid bilayers onto solid supports (Groves *et al.*, *Science*, **275:**651-653, 1997); *in situ* synthesis of oligonucleotides using a synthetic mask; as well as deposition of probes on porous sheets such as nitrocellulose sheets.

One method of immobilizing oligonucleotides onto a solid support is by the electrochemically directed synthesis of oligonucleotides on a solid support (Livache

et al., *Synthetic Metals*, 71:2143-2146, 1995). Briefly, this publication describes a manner of copolymerizing pyrrole, and pyrrole covalently linked to oligonucleotides via a spacer, giving rise to a solid copolymer film on the support formed by an oligonucleotide linked pyrrole chain. Such a construction has the disadvantage of a rapid loss of the stability of the immobilized polymeric film present on the electrodes.

C. Synthesis of High Density Array

Methods of forming high density arrays of oligonucleotides, peptides and other polymer sequences with a minimal number of synthetic steps are known. The oligonucleotide analog array can be synthesized on a solid substrate by a variety of methods, including, but not limited to, light-directed chemical coupling, and mechanically directed coupling. See Pirrung *et al.*, U.S. Patent No. 5,143,854 (see also PCT Application No. WO 90/15070) and Fodor *et al.*, PCT Publication Nos. WO 92/10092 and WO 93/09668 which disclose methods of forming vast arrays of peptides, oligonucleotides and other molecules using, for example, light-directed synthesis techniques. See also, Foder *et al.*, *Science*, **251**:767-77 (1991) These procedures for synthesis of polymer arrays are now referred to as VLSIPS™ procedures. using the VLSIPS™ approach, one heterogeneous array of polymers is converted, through simultaneous coupling at a number of reaction sites, into a different heterogeneous array. See, U.S. Patent Serial Nos. 07/796,243 and 07/980,523.

The development of VLSIPS™ technology as described in the above-noted U.S. Patent No. 5,143,854 and PCT Patent Publication Nos. WO 90/15070 and 92/10092.

In brief, the light-directed combinatorial synthesis of oligonucleotide arrays on a glass surface proceeds using automated phosphoramidite chemistry and chip masking techniques. In one specific implementation, a glass surface is derivatized with a silane reagent containing a functional group, e.g., a hydroxyl or amine group blocked by a photolabile protecting group. Photolysis through a photolithographic

mask is used selectively to expose functional groups which are then ready to react with incoming 5'-photoprotected nucleoside phosphoramidites. The phosphoramidites react only with those sites which are illuminated (and thus exposed by removal of the photolabile blocking group). Thus, the phosphoramidites only add to those areas selectively exposed from the preceding step. These steps are repeated until the desired array of sequences have been synthesized on the solid surface. Combinatorial synthesis of different oligonucleotide analogs at different locations on the array is determined by the pattern of illumination during synthesis and the order of addition of coupling reagents.

II. ARRAYS OF NUCLEIC ACIDS IN A WELL ARRANGEMENT

At times it is desired to form the arrays of nucleic acids utilizing a multi-well arrangement, for example, 256, 1024, well arrangements etc. well arrangement. Each well typically contains a small amount of fluid and one or more species of probes. While an array of nucleic acid probes present in a well arrangement obviously can contain a smaller number of species of nucleic acids than an array on a solid substrate (due to the relatively large space each well occupies) it nevertheless allows to carry out more complex reactions and chemical manipulations taking place in the liquid of the well and at times is advantageous.

III. LABELING OF NUCLEIC ACIDS

Various labeling methods are specified, for example in WO 97/27317. The nucleic acids which hybridized to the probes are detected by detecting one or more labels attached to the sample nucleic acids. The labels may be incorporated by any of a number of means well known to those of skill in the art. For example, the label can be simultaneously incorporated during the amplification step in the preparation of the sample nucleic acids. For example, if polymerase chain reaction (PCR) is carried out with labeled primers or labeled nucleotides a labeled amplification product will be available. The nucleic acid (e.g. DNA) is to be amplified in the

presence of labeled deoxynucleotide triphosphates (dNTPs). The amplified nucleic acid is then exposed to a nucleic acid array, and the extent of hybridization determined by the amount of label now associated with the array. In a preferred embodiment. transcription amplification, as described above, using a labeled nucleotide ((e.g. fluorescein- labeled UTP and/or CTP) incorporates a label into the transcribed nucleic acids.

Alternatively, a label may be added directly to the original nucleic acid sample (e.g., mRNA, polyA mRNA, cDNA, etc.) or to the amplification product after the amplification is completed. Such labeling can result in the increased yield of amplification products and reduce the time required for the amplification reaction. Means of attaching labels to nucleic acids include, for example, nick translation or end-labeling (e.g. with a labeled RNA) by kinasing of the nucleic acid and subsequent attachment (ligation) of a nucleic acid linker joining the sample nucleic acid to a label (e.g. a fluorophore).

Detectable labels suitable for use in the present invention include any composition detectable by spectroscopic, photochemical, biochemical, immunochemical, electrical, optical or chemical means. Useful labels in the present invention include biotin for staining with labeled streptavidin conjugate, magnetic beads (e.g. Dynabeads™), fluorescent dyes (e.g. fluorescein, texas red, rhodamine, green fluorescent protein, and the like, see, e.g. Molecular Probes, Eugene, Oregon, USA), radiolabels (e.g., ^{3}H, ^{125}I, ^{35}S, ^{14}C, or ^{32}P), enzymes (e.g. horse radish peroxidase, alkaline phosphatase and other commonly used in an ELISA), and colorimetric labels such as colloidal gold (e.g. gold particles in the 40-80 nm diameter size range scatter green light with high efficiency) or colored glass or plastic (e.g. polystyrene, polypropylene, latex, etc.) beads. Patents teaching the use of such labels include U.S. Patent Nos. 3,817,837; 3,850,752; 3,939,350; 3,996,345; 4,277,437; 4,275,149 and 4.366,241.

A fluorescent label is preferred because it provides a very strong signal with low background. It is also optically detectable at high resolution and sensitivity through a quick scanning procedure. The nucleic acid samples can all be labeled

with a single label, e.g. a single fluorescent label. Alternatively, in another embodiment, different nucleic acid samples can be simultaneously hybridized where each nucleic acid sample has a different label. For instance, one target could have a green fluorescent label and a second target could have a red fluorescent label. The scanning step will distinguish cites of binding of the red label from those binding the green fluorescent label. Each nucleic acid sample (target nucleic acid) can be analyzed independently from one another.

The label may be added to the target (sample) nucleic acid(s) prior to, or after the hybridization. So called *"direct labels"* are detectable labels that are directly attached to or incorporated into the target (sample) nucleic acid prior to hybridization. In contrast, so called *"indirect labels"* are joined to the hybrid duplex after hybridization. Often, the indirect label is attached to a binding moiety that has been attached to the target nucleic acid prior to the hybridization. Thus, for example, the target nucleic acid may be biotinylated before the hybridization. After hybridization, an avidin-conjugated fluorophore will bind the biotin bearing hybrid duplexes providing a label that is easily detected. For a detailed review of methods of labeling nucleic acids and detecting labeled hybridized nucleic acids see *Laboratory Techniques in Biochemistry and Molecular Biology*, Vol. 24: Hybridization with Nucleic Acid Probes, P. Tijssen, ed. Elsevier, N.Y., (1993).

IV. HYBRIDIZATION CONTROLS

A. Normalization controls

Normalization controls are nucleic acid probes that are perfectly complementary to labeled reference oligonucleotides that are added to the nucleic acid sample. The signals obtained from the normalization controls after hybridization provide a control for variations in hybridization conditions, label intensity, *"reading"* efficiency and other factors that may cause the signal of a perfect hybridization to vary between arrays. In a preferred embodiment, signals (e.g., fluorescence intensity) read from all other probes in the array are divided by

the signal (e.g.. fluorescence intensity) from the control probes thereby normalizing the measurements.

B. Mismatched controls

Mismatch controls may also be provided for the probes to the target genes, for expression level controls or for normalization controls. Mismatch controls are nucleic acid probes identical to their corresponding test or control probes except for the presence of one or more mismatched bases. A mismatched base is a base selected so that it is not complementary to the corresponding base in the target sequence to which the probe would otherwise specifically hybridize. One or more mismatches are selected such that under appropriate hybridization conditions (e.g. stringent conditions) the test or control probe would be expected to hybridize with its target sequence, but the mismatch probe would not hybridize (or would hybridize to a significantly lesser extent). Preferred mismatch probes contain a central mismatch. Thus. for example, where a probe is a 20 mer, a corresponding match probe will have the identical sequence except for a single base mismatch (e.g., substituting a G, a C or a T for an A) at any of positions 6 through 14 (the central mismatch).

V. SAMPLE PREPARATION

In the simplest embodiment, a nucleic acid sample is the total mRNA or a total cDNA isolated and/or otherwise derived from a biological sample. The term *"biological sample"*, as used herein, refers to a sample obtained from an organism or from components (e.g.. cells) of an organism. The sample may be of any biological tissue or fluid. Frequently the sample will be a *"clinical sample"* which is a sample derived from a patient. Such samples include, but are not limited to, sputum. blood, blood cells (e.g. white cells). tissue or fine needle biopsy samples, urine, peritoneal fluid, and pleural fluid, or cells therefrom. Biological samples may also include sections of tissue such as frozen sections taken for histological purposes.

The nucleic acid (either genomic DNA or mRNA) may be isolated from the sample according to any of a number of methods well known to those of skill in the art. One of skill will appreciate that where alterations in the copy number of a gene are to be detected genomic DNA is preferably isolated. Conversely, where expression levels of a gene or genes are to be detected, preferably RNA (mRNA) is isolated.

Methods of isolating total mRNA are well known to those of skill in the art. For example, methods of isolation and purification of nucleic acids are described in detail in Chapter 3 of Laboratory Techniques in *Biochemistry and Molecular Biology: Hybridization with Nucleic Acid Probes*, Part I. *Theory and Nucleic Acid Preparation*, P. Tijssen, ed. Elsevier, N.Y. (1993).

In a preferred embodiment, the total nucleic acid is isolated from a given sample using, for example, an acid guanidinium-phenol-chloroform extraction method and polyA$^+$ mRNA is isolated by oligo dT column chromatography or by using (dT)n magnetic beads (see, e.g. Sambrook *et al., Molecular Cloning: A Laboratory Manual* (2nd ed.), Vols. 1-3, Cold Spring Harbor Laboratory, (1989), or *Current Protocols in Molecular Biology*, F. Ausubel *et al.*, ed. Greene Publishing and Wiley-Interscience, New York (1987)).

Frequently, it is desirable to amplify the nucleic acid sample prior to hybridization. One of skill in the art will appreciate that whatever amplification method is used, if a quantitative result is desired, case must be taken to use a method that maintains or controls for the relative frequencies of the amplified nucleic acids.

Other suitable amplification methods include, but are not limited to polymerase chain reaction (PCR) (Innis, *et al., PCR Protocols: A Guide to Methods and Application*, Academic Press, Inc. San Diego, (1990)), ligase chain reaction (LCR) (see Wu and Wallace, *Genomics*, **4**:460 (1989), Landegren *et al. Science*, **241**:1077 (1988) and Barringer, *et al., Gene*, **89**:117 (1990), transcription amplification (Kwoh, *et al., Proc. Natl. Acad. Sci. USA*, **86**:1173 (1989), and

self-sustained sequence replication (Guatelli, *et al.*, *Proc. Nat. Acad. Sci. USA*, 87:1874 (1990)).

VI. HYBRIDIZATION BETWEEN NUCLEIC ACIDS IN THE SAMPLE AND THE ARRAY

Nucleic acid hybridization simply involves providing a denatured probe and target nucleic acid under conditions where the probe and its complementary target can form stable hybrid duplexes through complementary base pairing. The nucleic acids that do not form hybrid duplexes are then washed away leaving the hybridized nucleic acids to be detected, typically through detection of an attached detectable label. It is generally recognized that nucleic acids are denatured by increasing the temperature or decreasing the salt concentration of the buffer containing the nucleic acids, or in the addition of chemical agents, or the raising of the pH. Under low stringency conditions (e.g. low temperature and/or high salt and/or high target concentration) hybrid duplexes (e.g., DNA:DNA, RNA:RNA, or RNA:DNA) will form even where the annealed sequences are not perfectly complementary. Thus specificity of hybridization is reduced at lower stringency. Conversely, at higher stringency (e.g. higher temperature or lower salt) successful hybridization requires fewer mismatches.

VII. DETECTION METHODS

Methods for detection depend upon the label selected and are known to those of skill in the art. Thus, for example, where a colorimetric label is used, simple visualization of the label is sufficient. Where a radioactive labeled probe is used, detection of the radiation (e.g. with photographic film or a solid state detector) is sufficient.

As explained above, the use of a fluorescent label is preferred because of its extreme sensitivity and simplicity. Standard procedures are used to determine the positions where interactions between a target sequence and a reagent take place.

For example, if a target sequence is labeled and exposed to an array of different oligonucleotide probes, only those locations where the oligonucleotides interact with the target (sample nucleic acid(s)) will exhibit significant signal. In addition to using a label, other methods may be used to scan the matrix to determine where interaction takes place. The spectrum of interactions can, of course, be determined in a temporal manner by repeated scans of interactions which occur at each of a multiplicity of conditions. However, instead of testing each individual interaction separately, a multiplicity of sequence interactions may be simultaneously determined on the array.

In a preferred embodiment, the hybridized array is excited with a light source at the excitation wavelength of the particular fluorescent label and the resulting fluorescence at the emission wavelength is detected. In a particularly preferred embodiment, the excitation light source is a laser appropriate for the excitation of the fluorescent label.

Detection of the fluorescence signal preferably utilizes a confocal microscope, more preferably a confocal microscope automated with a computer-controlled stage to automatically scan the entire high density array. The microscope may be equipped with a phototransducer (e.g. a photomultiplier, a solid state array, a ccd camera, etc.) attached to an automated data acquisition system to automatically record the fluorescence signal produced by hybridization to each oligonucleotide probe on the array. Such automated systems are described at length in U.S. Patent No.: 5,143,854, PCT Application 20 92/10092, and copending U.S.S.N. 08/195,889 filed on February 10, 1994. Use of laser illumination in conjunction with automated confocal microscopy for signal detection permits detection at a resolution of better than about 100 μm, more preferably better than about 50 μm, and most preferably better than about 25 μm.

VIII. ANALYSIS OF DETECTION RESULTS

One of skill in the art will appreciate that methods for evaluating the hybridization results vary with the nature of the specific probe nucleic acids used as

well as the controls provided. In the simplest embodiment, simple quantification of the fluorescence intensity for each probe is determined. This is accomplished simply by measuring probe signal strength at each location (representing a different probe) on the high density array (e.g. where the label is a fluorescent label, detection of the amount of fluorescence (intensity) produced by a fixed excitation illumination at each location on the array). Comparison of the absolute intensities of an array hybridized to nucleic acids from a *"test"* sample with intensities produced by a *"control"* sample provides a measure of the relative abundance of the nucleic acids that hybridize to each of the probes.

One of skill in the art, however, will appreciate that hybridization signals will vary in strength with efficiency of hybridization, the amount of label on the sample nucleic acid and the amount of the particular nucleic acid in the sample. Typically nucleic acids present at very low levels (e.g., < 1 pM) will show a very weak signal. At some low level of concentration, the signal becomes virtually indistinguishable from the background. In evaluating the hybridization data, a threshold intensity value may be selected below which a signal is not counted as being essentially indistinguishable from background.

Where it is desirable to detect nucleic acids expressed at lower levels, a lower threshold is chosen. Conversely, where only high expression levels are to be evaluated a higher threshold level is selected. In a preferred embodiment, a suitable threshold is about 10% above that of the average background signal.

In addition, the provision of appropriate controls permits a more detailed analysis that controls for variations in hybridization conditions, cell health, non-specific binding and the like. Thus, for example, in a preferred embodiment, the hybridization array is provided with normalization controls as described above. These normalization controls are probes complementary to control sequences added in a known concentration to the sample. Where the overall hybridization conditions are poor, the normalization controls will show a smaller signal reflecting reduced hybridization. Conversely, where hybridization conditions are good, the normalization controls will provide a higher signal reflecting the improved

hybridization. Normalization of the signal derived from other probes in the array to the normalization controls thus provides a control for variation in array synthesis or in hybridization conditions. Typically, normalization is accomplished by dividing the measured signal from the other probes in the array by the average signal produced by the normalization controls. Normalization may also include correction for variations due to sample preparation and amplification. Such normalization may be accomplished by dividing the measured signal by the average signal from the sample preparation/amplification control probes (e.g., the *BioB* probes). The resulting values may be multiplied by a constant value to scale the results.

As indicated above, the high density array can include mismatch controls or, in the case of generic difference screening arrays, pairs of related oligonucleotide probes differing in one or more preselected nucleotides. In preferred expression monitoring arrays, there is a mismatch control having a central mismatch for every probe (except the normalization controls) in the array. It is expected that after washing in stringent conditions, where a perfect match would be expected to hybridize to the probe, but not to the mismatch, the signal from the mismatch controls should primarily reflect non-specific binding or the presence in the sample of a nucleic acid that hybridizes with the mismatch. In expression monitoring analyses, where both the probe in question and its corresponding mismatch control both show high signals, or the mismatch shows a higher signal than its corresponding test probe, the signal from those probes is preferably ignored. The difference in hybridization signal intensity between the target specific probe and its corresponding mismatch control is a measure of the discrimination of the target-specific probe. Thus, in a preferred embodiment, the signal of the mismatch probe is subtracted from the signal from its corresponding test probe to provide a measure of the signal due to specific binding of the test probe. Similar, as discussed below, in generic difference screening, the difference between probe pairs is calculated.

The concentration of a particular sequence can then be determined by measuring the signal intensity of each of the probes that bind specifically to that nucleic acid and normalizing to the normalization controls. Where the signal from

the probes is greater than the mismatch, the mismatch is subtracted. Where the mismatch intensity is equal to or greater than its corresponding test probe, the signal is ignored. The expression level of a particular gene can then be scored by the number of positive signals (either absolute or above a threshold value), the intensity of the positive signals (either absolute or above a selected threshold value), or a combination of both metrics (e.g., a weighted average).

When performing the assay of the invention it is possible to work under physical/chemical conditions which by themselves will give information on binding affinity of targets to probes, such as for example changing the ionic strength or the temperature, performing a gradual rinsing to gradually remove first targets with a relatively weak interaction with the probes and subsequently targets with a higher degree of interaction and using this information in constructing of vector c. As can be understood the present invention is free from the constraints of prior art assay methods that require that the melting point of all target-probe hybrids be about the same.

IX. ILLUSTRATION OF THE MANNER OF CARRYING OUT THE INVENTION

In accordance with the present invention as shown in Fig. 1, a solid support **2** onto which a probe array, generally designated by **4**, has been synthesized, is produced. Using the notation and terminology introduced above, the probe array **4** comprises a total of k different probing units, each of which in this specific embodiment consists of a single probe oligonucleotide species, five of which are shown as $\mathbf{P_1}$-$\mathbf{P_5}$. The nucleotide sequence of the probe at each location in the array is known. A sample **8** is prepared containing n different labeled target species five of which are shown as $\mathbf{S_1}$-$\mathbf{S_5}$. The number of probe species (k) may not be equal to the number of target species (n). A given target species in the target mixture **8** may bind to more than one probe species in the probe array **4**. The matrix **T** is defined as above. **T** may be non-square and may have rows or columns containing each more than one non-zero element.

As shown in Fig. 2, the probe array **4** is incubated in the presence of sample **8** under conditions allowing the labeled probe targets in the target mixture **8** to hybridize with probes in the probe array. As depicted in Fig. 2, target species $\mathbf{T_1}$ has bound to probe species $\mathbf{P_2}$ and $\mathbf{P_4}$, while target species $\mathbf{S_3}$ has hybridized with, probe species $\mathbf{P_2}$, $\mathbf{P_3}$, and $\mathbf{P_5}$. Accordingly, $t_{21} = t_{41} = t_{23} = t_{33} = t_{53} = 1$.

After the incubation, unbound targets are remove as shown in Fig. 3. The amount of label associated with each probe species is measured so as to provide the k-dimensional vector **c** of measured label.

The n-dimensional vector **e** of the relative abundance of the target species in the original target mixture is related to **T** and **b** by Equation **(1)**. The vector e is obtained by solving Equation (1).

An illustrative, simmulated example of a case wherein the assayed sample is assayed for the presence therein of one of 9 targets - S_0, S_1, S_2, S_3, S_4, S_5, S_6, S_7 and S_8. As will be shown the assay of these targets may be performed by a total of 7 probes – P_1, P_2, P_3, P_4, P_5, P_6 and P_7. The assumption is that the vector is scarce meaning here that in each sample no more than 2 targets are exprssed.

The probes $P_1 \ldots P_7$ are constructed to have target specificity as illustrated in the following Table 1:

Table 1

Serial Number	Target	Probe
00	S_1	P_1, P_4, P_7
01	S_2	P_1, P_5
02	S_3	P_1, P_6
10	S_4	P_2, P_4
11	S_5	P_2, P_5, P_7
12	S_6	P_2, P_6
20	S_7	P_3, P_4
21	S_8	P_3, P_5
22	S_9	P_3, P_6, P_7

The probes are constructed so that they will bind the targets as represented in the table. The serial number of the target in the Table is given in base 4 as it conveniently translates in this case into the manner of constructing the appropriate probes: where the left digit is 0, 1 or 2, this means that the respective target binds to probe S_1, S_2 or S_3, respectively; where the right digit is 0, 1 or 2, this means that the respective target binds to probe S_4, S_5 or S_6, respectively; where the two digits constituting the serial number are the same, this means that these targets bind also to P_7.

In the same manner probes for different number of targets may be constructed.

Probe P_1, for example may be constructed to include a sequence from S_0, S_1 and S_2; probe P_2 to include a sequence from S_3, S_4 and S_5; etc.

Assume for example the case where P_1, P_2, P_4 and P_5 "light-up" (namely indicate that a target has hybridized thereto). The only solution is that S_1 and S_3 are in the sample (as the assumption as noted above is that there are no more than two targets in the assayed sample). If probe P_7 would also "light-up" than this would mean that targets S_0 and S_4 are in the assayed sample. If, for example only two probes - P_1 and P_5, or three probes - P_1, P_4 and P_7 "light-up" this would mean the existence of only one target in the assayed sample – S_1 and S_0 in this specific example.